Machine Learning in the audio domain

When the neural network is overkill or where are the limits of lightweight models

Co-organizer

Roman Smirnov

Machine Learning Engineer

- 1 year at Exness
- 7 years as an MLE/Tech Lead at Skyeng & MSU Labs

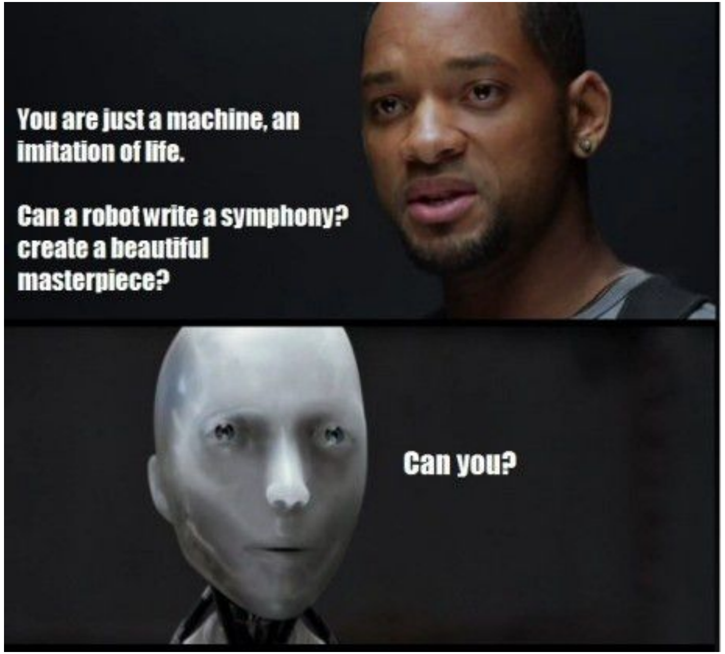
Co-organizer

Table of contents

Audio domain

Analyzing sounds using neural networks

- Call Centers
- Virtual Assistants
- Speech and music generation



Pic source
https://www.linkedin.com/pulse/turing-test-music-why-spotify-flooded-songs-mathieu-govaarts/

Problem Statement

Data modality and SotA

When working with media data, we usually use large neural networks.

But:

- They are resource-heavy
- They are slower than light models

Usually it is **Transformers models**

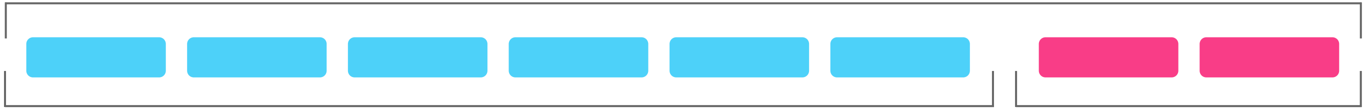
Pic source https://comicvine.gamespot.com/

Time for SotA

It takes **LONG** time to train SotA model for the task we'll discuss today

Train

Inference



15h for first epoch, 40k audio files

7h to read and prepare audio files

1.3s for one file processing (including reading)

GPU: RTX 3090Ti, 24GB

Audio-domain tasks

Understanding and Generation

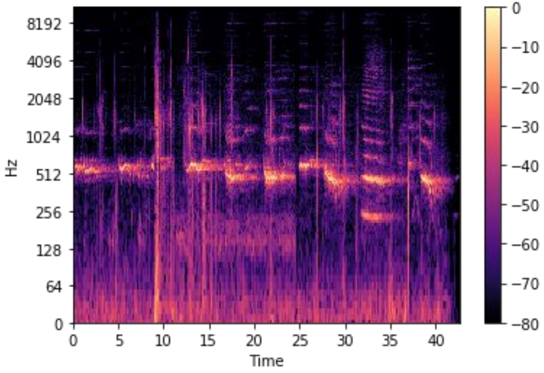
Understanding (Classification)

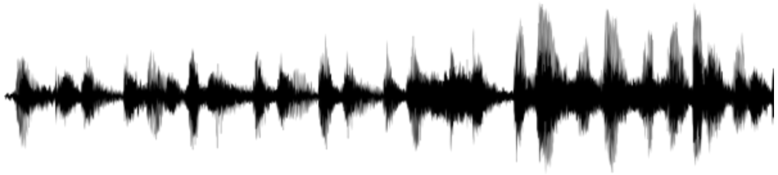
- Classification, e.g. emotions classification
- “Token classification”—voice activity detection
- Speaker separation, user verification
- Automatic Speech Recognition (ASR)

Pic source https://kairosgame.com/

Generation

- Voice cloning
- Text to speech
- Speech and music generation





Pic source https://twitter.com/ai__pub https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0
https://www.holdcom.com/support/telephony-audio-file-formats/

Experiments & Business

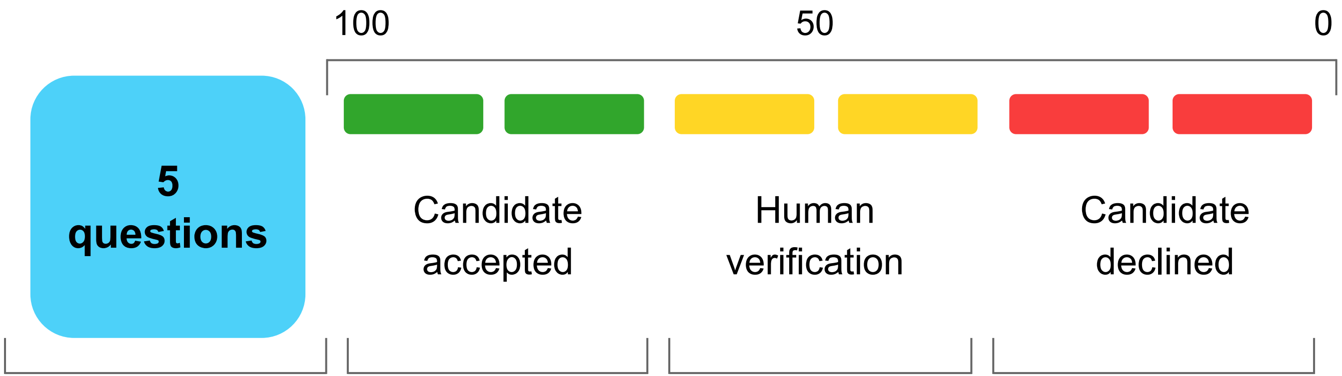
Classification

Evaluation of the call quality for the call center:

Stable connection	?	Unstable connection

Regression

Interviewing a candidate for an English Teacher position

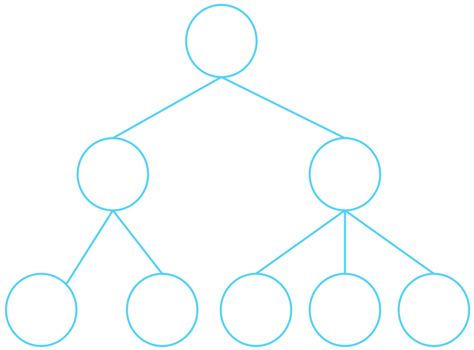


Fluency and pronunciation are evaluated separately using 5-points-scale

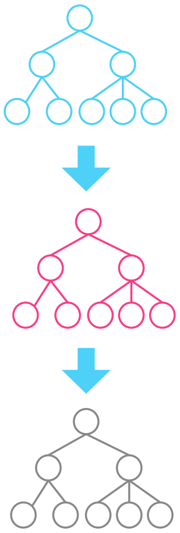
Gradient Boosting on Decision Trees (GB on DT)

- Fast
- Work on statistics aggregates of audio
- But is it accurate?

Single Decision Tree

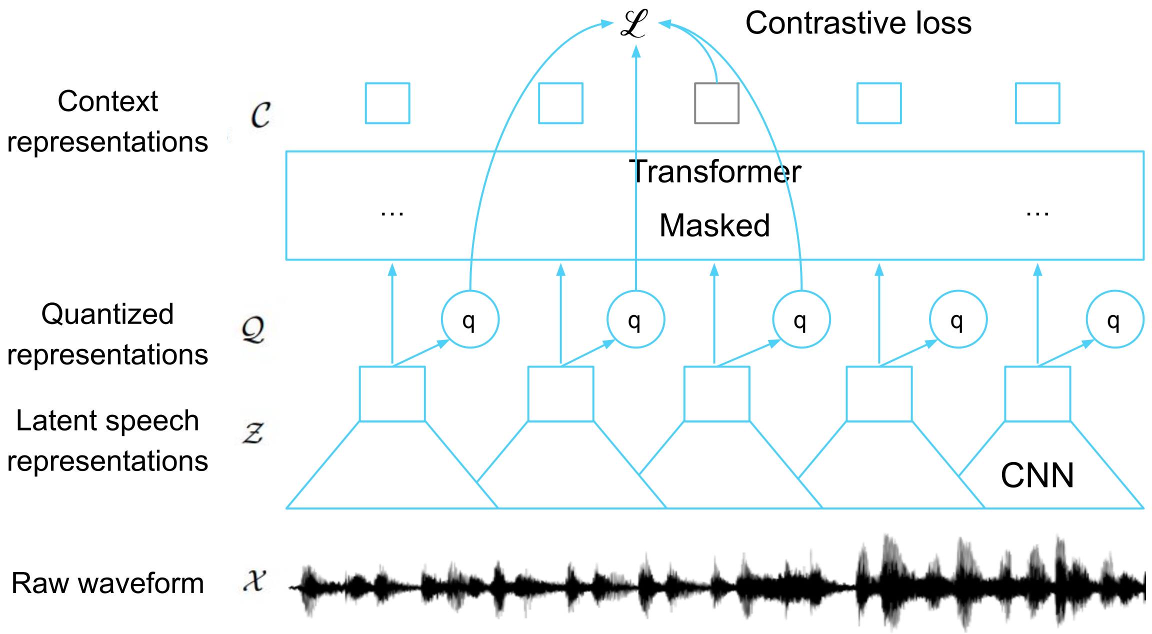


Gradient Boosted Trees



Wav2Vec2

- SotA for speech recognition
- Transformer model
- CNNs allows to encode local features
- But is it fast?



Training Process

Preprocessing

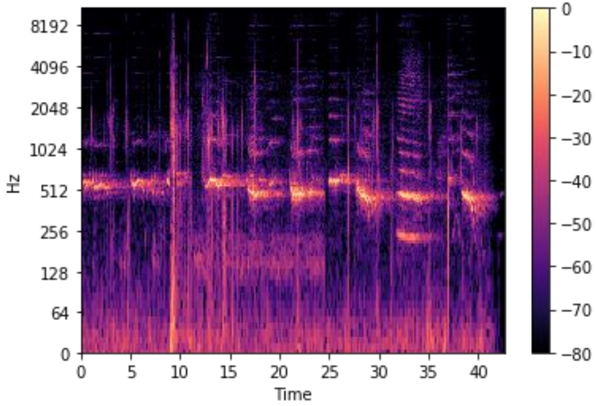
For GB on DT

- Collect amplitude and melspectrograms statistics
- Mean, median, min, max, std, kurtosis, skew… over mel for every frequency and raw wav

For Wav2Vec2

- Resample to 16kHz
- Just truncation

What is melspectrogram?

Raw audio	Melspectrogram
	

Training

For GB on DT

- Catboost
- Small trees (depth 6)
- 100-10000 iterations

For Wav2Vec2

- Wav2Vec2 base
- Freeze feature encoder
- Learning rate schedulers
- 10 epochs

Time metrics. Classification

For GB on DT

- Train: 45 min to read data
- 0.5s to train on GPU
- 5k audio-files
- Inference: 5 files, 2s to read data, 0.1s to inference on CPU

For Wav2Vec2

- Train: 45 min to read data
- 5h to train on GPU (10 epochs)
- 5k audio-files, batch size 6
- Inference: 5 files, 2s to read data, 105s to inference on CPU

Time metrics. Regression

For GB on DT

- Train: 7h to read data
- 5 min to train on GPU
- 40k audio-files
- Inference: 5 files, 2s to read data, 0.1s to inference on CPU

For Wav2Vec2

- Train: 7h to read data
- 90h to train on GPU (10 epochs)
- 40k audio-files, batch size 6
- Inference: 5 files, 2s to read data, 105s to inference on CPU

Results

Results. Classification task

Natural imbalance!

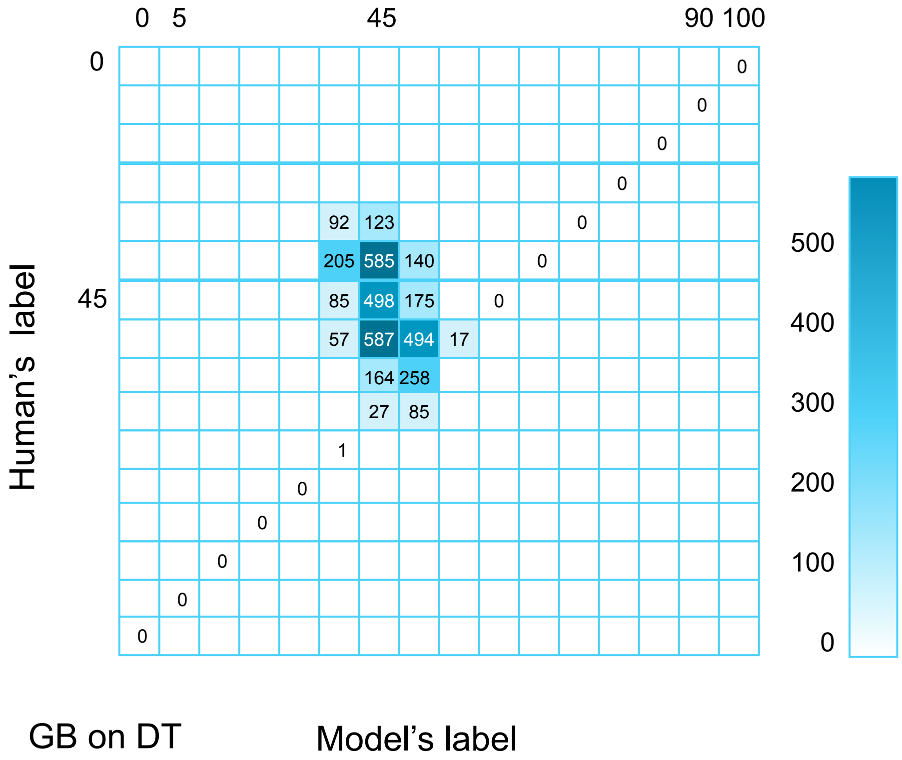
GB on DT

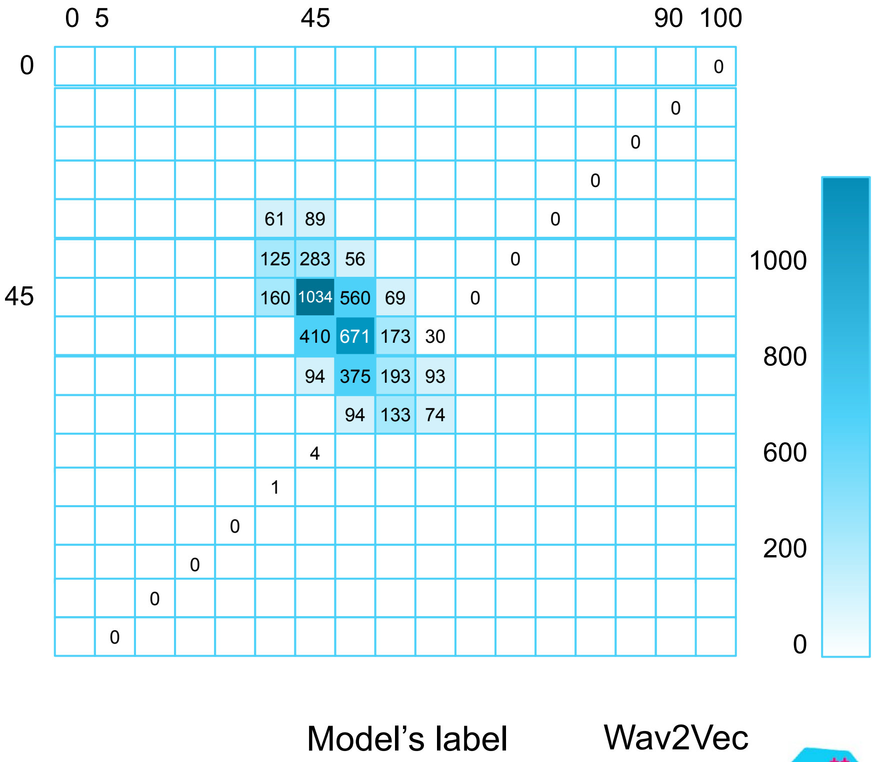
Humans label \ Models label	Slightly bad	Bad	Ok
Slightly bad	0	0	4
Bad	0	10	4
Ok	0	2	538

Wav2Vec

Humans label \ Models label	Slightly bad	Bad	Ok
Slightly bad	3	0	1
Bad	0	13	1
Ok	1	4	535

Results. Regression task





Business Outcomes

Outcomes

Time and $ / accuracy trade-off

- We took GB on DT for classification
- We took Wav2Vec2 for regression

Where are the limits and what is overkill?

Pic source https://medium.com/@devsociety_

Thank you for your attention!

Vote for my talk

[Roman Smirnov]

[rgsmirnov]

[Exness Tech Blog]

++ High Load Armenia

Co-organizer

Yandex